Part C - Mini Research Questions

Why Attention Beats RNNs/LSTMs: A Paradigm Shift in Sequence Modeling

1. Introduction

For decades, RNNs and their gated variants, Long Short-Term Memory (LSTM) networks, were the gold standard for sequence modeling. However, the 2017 seminal paper "Attention is All You Need" by Vaswani et al introduced the Transformer architecture, which fundamentally replaced recurrence with a self-attention mechanism. This analysis explores the three primary reasons for this dominance. parallelization, long-range dependency handling and computational efficiency

2. The Bottleneck of Recurrence.

The core limitation of RNNs/LSTMs lies in their sequential nature. To process a token at time step t, the model first compute the hidden state h_{t-1} from the previous step.

RNN Hidden State Equation

$$h_t = \sigma(W_{hh} h_{t-1} + W_{xh} x_t + b)$$

This linear dependency on time prevents parallelization during training. In contrast

The Transformer allows the model to process all tokens in a sequence simultaneously by calculating a "score" for how every token relates to every other token in one matrix operation.

3. The Vanishing Gradient & Long-Range Dependencies.

LSTMs were designed to mitigate the "vanishing gradient" problem using gates (input, forget, and output). However, even with these gates, information must still travel through a chain of operations. As the distance between two related words (eg. a subject and a distant verb) increases, the signal weakens.

LSTM Forget Gate Equation

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

In a Transformer, the "path length" between any two tokens is always O(1). Because self-attention relates every position to every other position directly the distance in the sequence does not degrade the signal.

4 The Self-Attention Mechanism

The "Attention" mechanism operates by mapping a query (Q), a key (K) and value (V) for each token.

PAGE No.

DATE / /

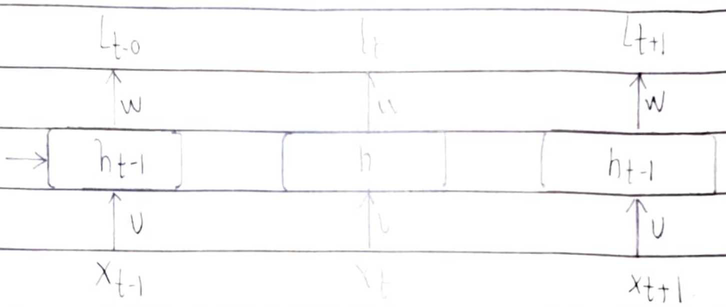
Scaled Dot-Product Attention Equation:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

This formula allows the model to build a dynamic "context vector" that focuses on relevant parts of the sequence regardless of their position.

5.

RNN structure.



6. Computation Complexity.

Feature	RNN/LSTM	Transformer
Complexity per layer	$O(nd^2$	$O(n^2 \cdot d)$.
Sequential Operations	$O(n)$	$O(1)$.
Maximum Path Length	$O(n)$	$O(1)$

Where n is sequence length and d is representation dimension.

Citations.

- Vaswani, A. , et al (2017) Attetion is All you need
- Hrochreiter S. & Schimidhuber J. (1997). Long Short Term Memory. Neural Computation.
- Bahdanau, D. et al (2014) Neural Machine Translation by Jointly Learning to Align and Translate . arXiv.